FIG. 1A
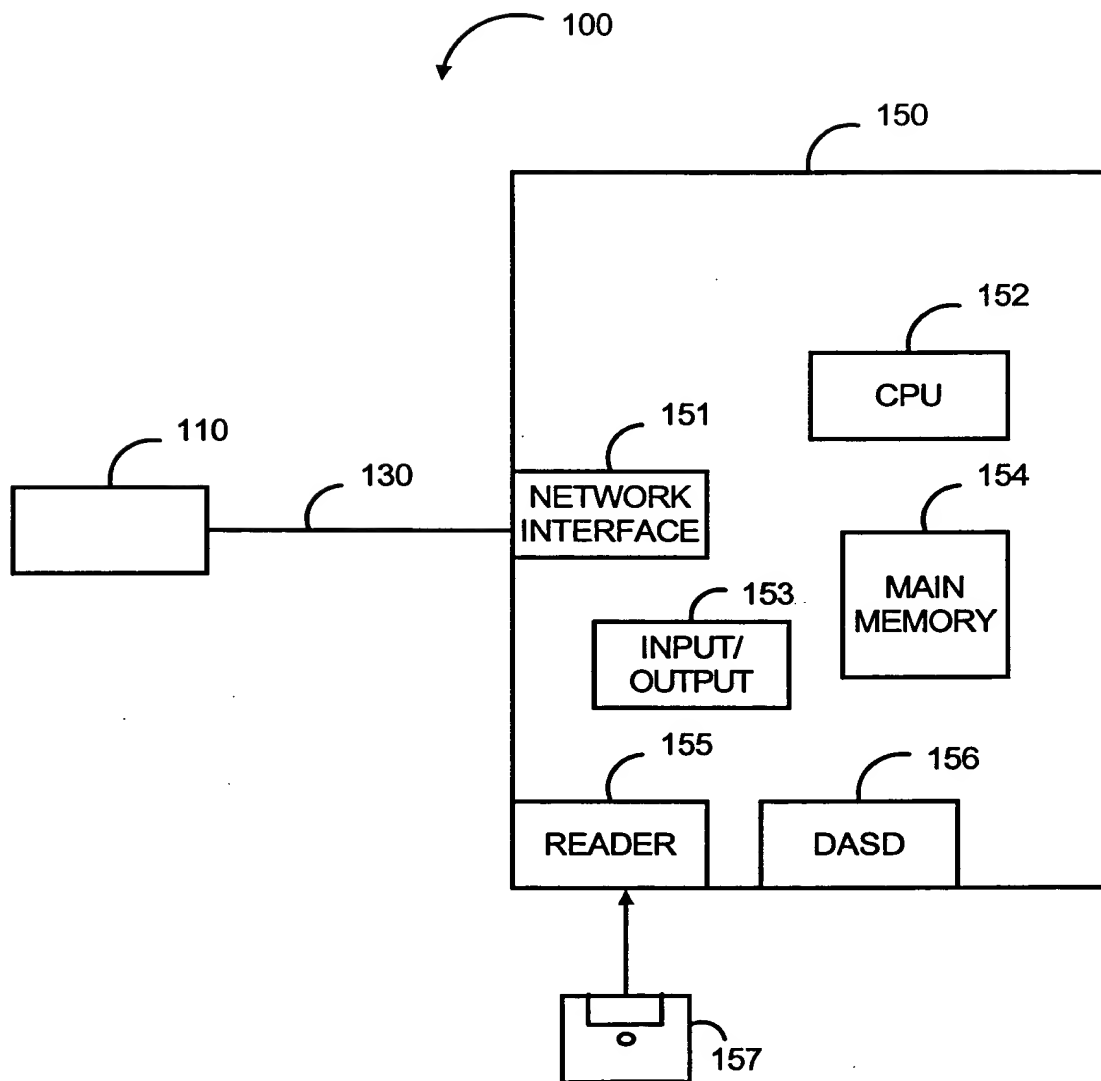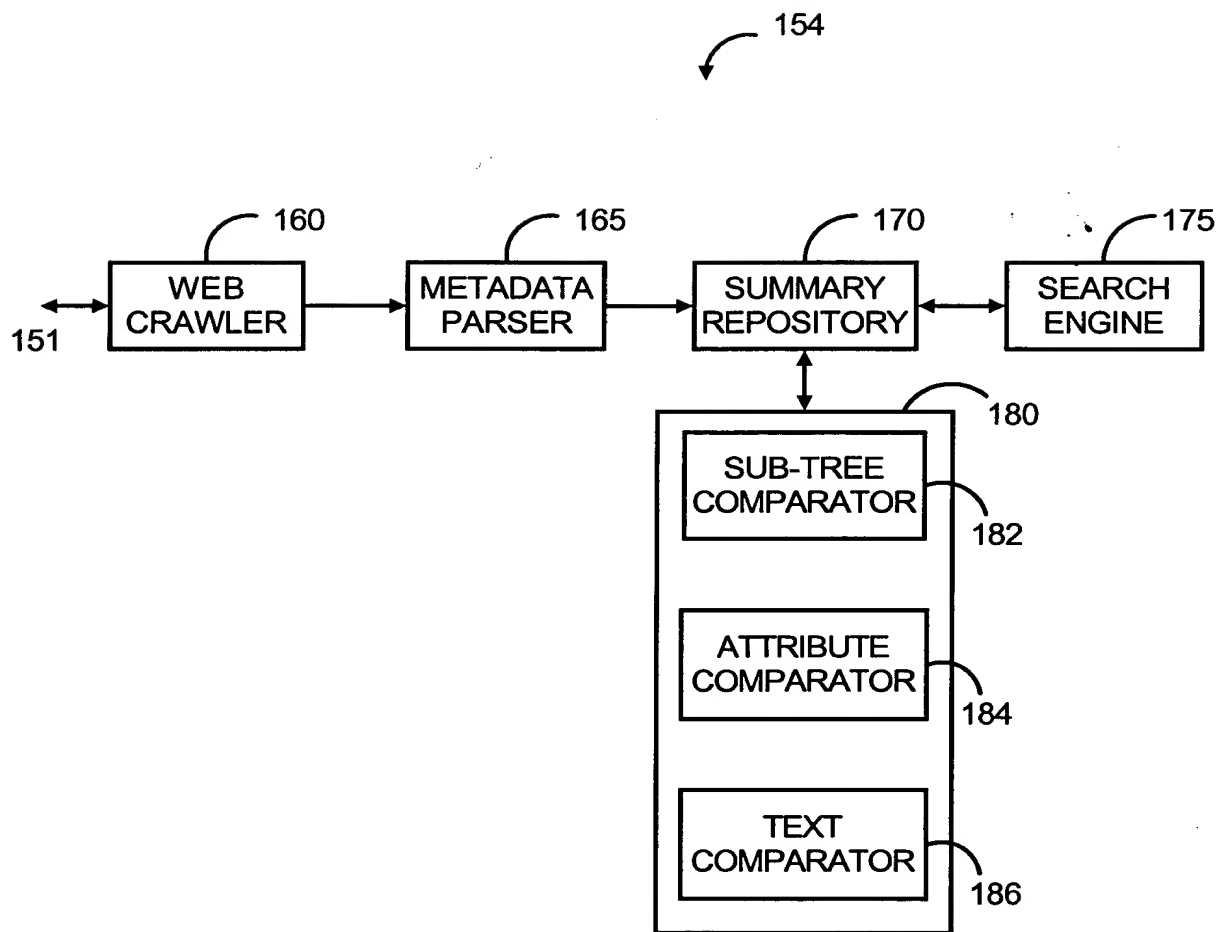
FIG. 1B

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/schemas/rdf-schema">
        <!--
                This RDF Description contains information about (1) the data gatherer and
                (2) the metadata of the data source.
        -->
        <rdf:Description  gatherer="Grand Central Station Gatherer II"
                          gathered-on="Tue Mar 23 17:38:40 GMT 1999"
                          summarizer="com.ibm.almaden.gcs.summarizer.HTMLSummaryMaker"

                          resource="http://www.people.com/jane_doe/homepage.html"
                          source-last-modified="Tue Feb 17 15:43:58 GMT 1999"
                          mime-type="http/html"
                          content-length="46220"
                          source-is="Good"
                          comments="good"/>
        <!--
                This RDF Description contains information about from the data source itself.
                In addition to textual information, it contains structural information, for example
                the URLs that the page points to.
        -->
        <rdf:Description  html-title="Jane's Homepage"          - 231
                          html-encoding="8859_1"                - 232
                          abstract="Personal homepage of Jane Doe">  - 233
                <!--
                        The "ref-annotations" contains summaries of the out-links of the HTML
                        page. In this case, the attribute "ref" gives the URL of the referenced
                        page. The attribute "annotation" gives the text associated with each
                        out-link.
                -->
                <ref-annotations>
                        <rdf:Bag>
                                <rdf:LI>
                                        <rdf:Description                                234a
                                                ref="http://www.yahoo.com/"
                                                annotation="Yahoo!"/>
                                </rdf:LI>
                                <rdf:LI>                                                234b              234c
                                        <rdf:Description
                                                ref="http://www.people.com/jane_doe/my_photo.jpg"
                                                annotation="picture of me."/>
                                </rdf:LI>                                                 234d
                        </rdf:Bag>
                </ref-annotations>
                <!--
                        The "presentation-text" contains the textual content of the HTML page
                        that may be seen through a WWW browser.
                -->
                <presentation-text>
                        <rdf:Bag>
                                <rdf:LI>Welcome to my homepage. </rdf:LI>
                                <rdf:LI>Use Yahoo! to search for something or look
                                        at a picture of me.</rdf:LI>
                        </rdf:Bag>
                </presentation-text>
        </rdf:Description>
</rdf:RDF>
```
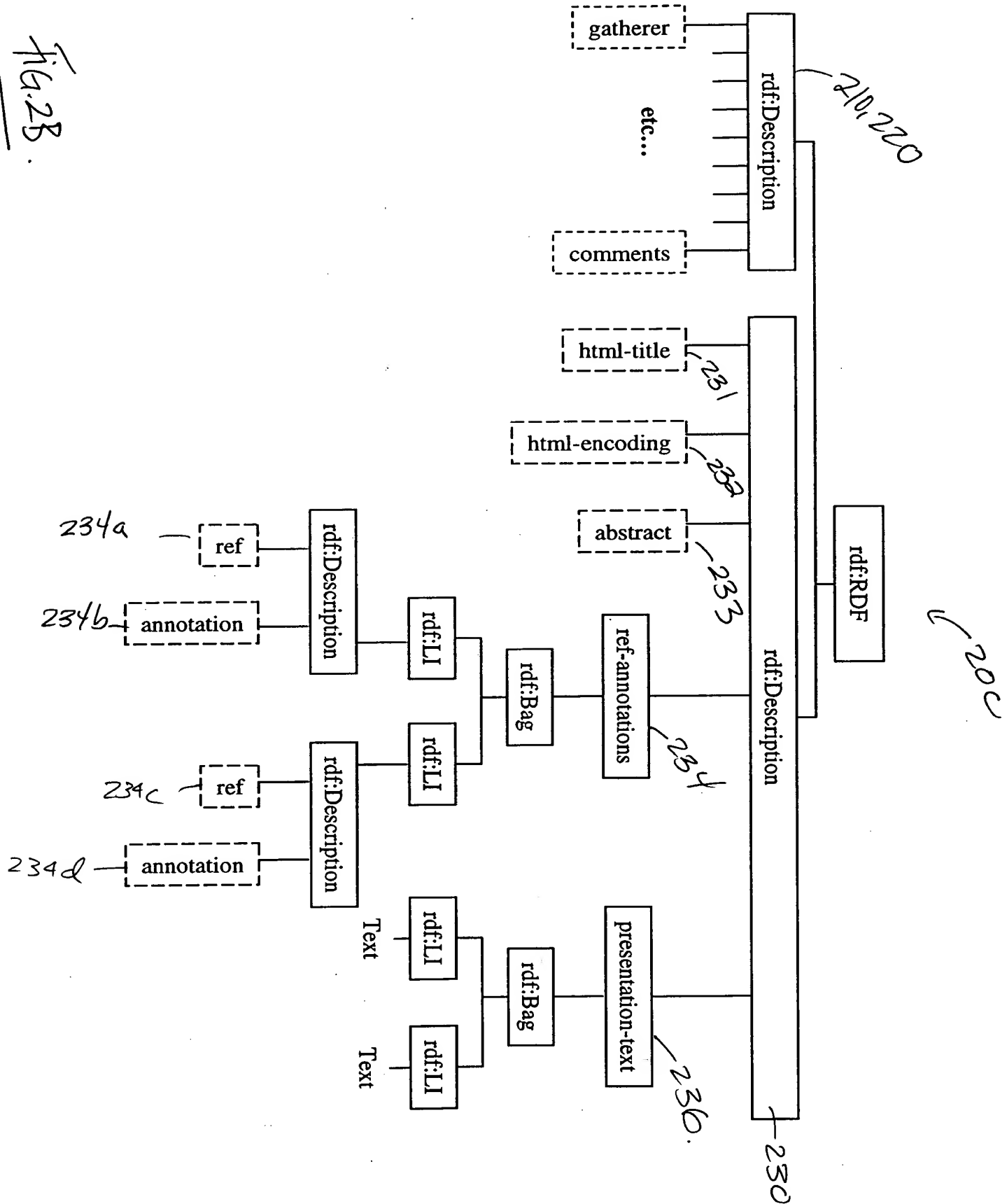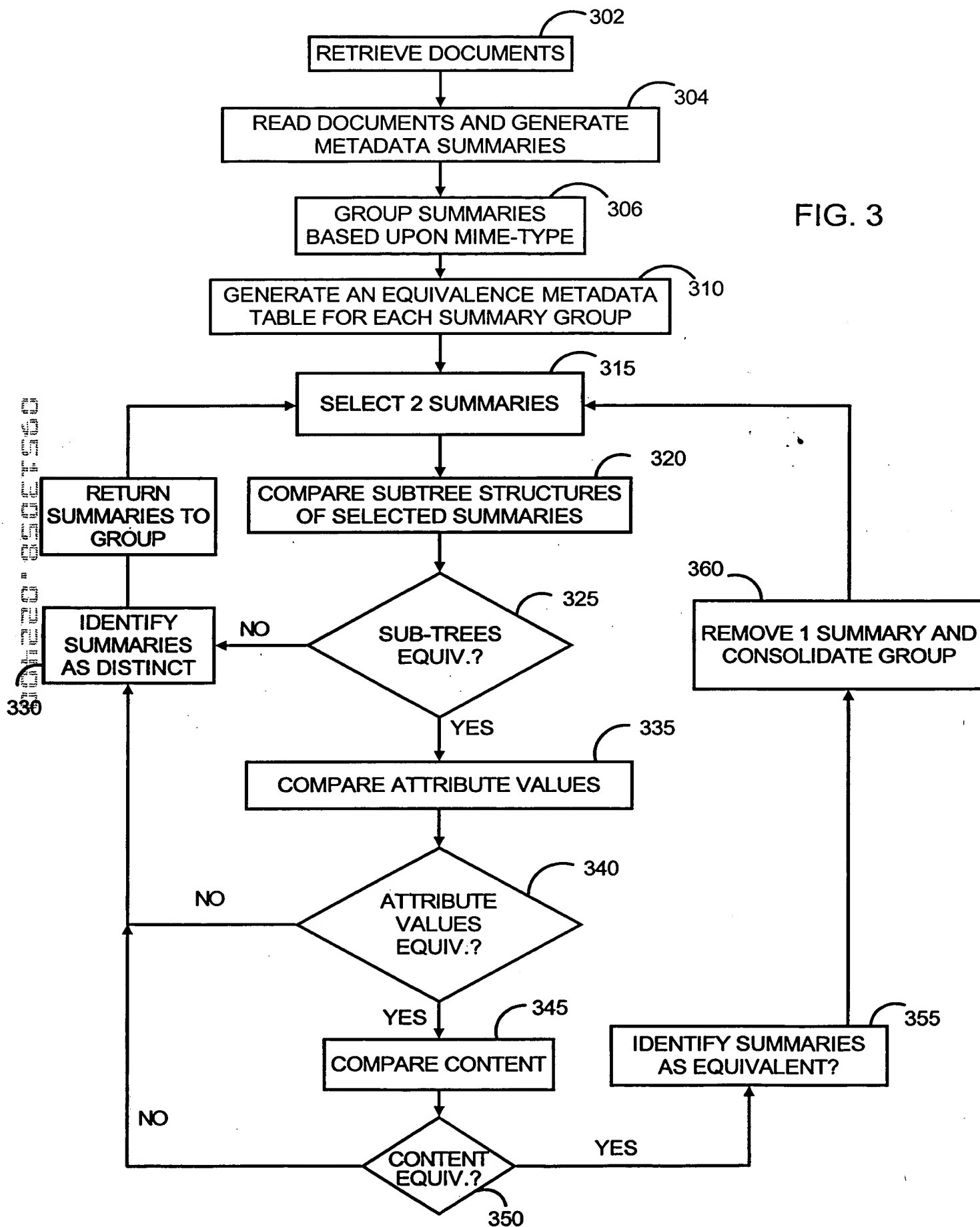
210

220

230

234

236

fig.2B.

gatherer

etc...

comments

rdf:Description

210, 220

html-title — 231

html-encoding — 232

abstract — 233

rdf:RDF

20C

234a — ref — rdf:Description

234b — annotation

rdf:LI

rdf:Bag

ref-annotations — 234

rdf:Description

234c — ref — rdf:Description

234d — annotation

rdf:LI

rdf:LI

Text — rdf:LI

rdf:Bag

presentation-text — 236.

Text — rdf:LI

230

FIG. 3

302 RETRIEVE DOCUMENTS

304 READ DOCUMENTS AND GENERATE METADATA SUMMARIES

306 GROUP SUMMARIES BASED UPON MIME-TYPE

310 GENERATE AN EQUIVALENCE METADATA TABLE FOR EACH SUMMARY GROUP

315 SELECT 2 SUMMARIES

RETURN SUMMARIES TO GROUP

320 COMPARE SUBTREE STRUCTURES OF SELECTED SUMMARIES

330 IDENTIFY SUMMARIES AS DISTINCT ← NO — 325 SUB-TREES EQUIV.?

360 REMOVE 1 SUMMARY AND CONSOLIDATE GROUP

YES

335 COMPARE ATTRIBUTE VALUES

NO ← 340 ATTRIBUTE VALUES EQUIV.?

YES 345 COMPARE CONTENT

355 IDENTIFY SUMMARIES AS EQUIVALENT?

NO

350 CONTENT EQUIV.? — YES →

315

315a

ORDER METADATA
SUMMARIES INTO X
GROUPS
$F = (f_1, f_2, f_3, \ldots f_n)$
$G = (g_1, g_2, g_3, \ldots g_n)$

315b

SELECT m^{th} GROUP

315c

SELECT i^{th} REFERENCE SUMMARY
AND SUBGRAPH

315d

SELECT j^{th} SECONDARY
SUMMARY AND SUBGRAPH; $j > i$

COMPARE i^{th} AND j^{th} SUMMARIES PER
STEPS 320-360

315e

NO

$j = n$
?

INCREM. j

315f

YES

315g

NO

$i = n-1?$

INCREM. i

315h

YES

315i

NO

$m = x$
?

YES

INCREM. m

END

315k

FIG. 4